Educational Research Bulletin

NOVEMBER 19, 1930

VOL. IX, No. 17

Measuring the Ability to Infer

By RALPH W. TYLER

THE necessity of having objective evidence as to the degree to which students are attaining each of the important objectives of education is stressed in an editorial in this issue of the BULLETIN. This need has been recognized in several departments of Ohio State University. Instructors in these departments in co-operation with the writer are attacking the problem of developing methods for measuring the students' attainment of these objectives which are not usually tested in the ordinary examination. The work in the Department of Zoölogy may serve to illustrate the general program now under way in several other departments.

As a result of considerable deliberation, the instructors in zoölogy have set the following goals toward which students should progress in elementary courses:

1. A fund of information about animal activities and structures
2. An understanding of technical terminology
3. An ability to draw inferences from facts, that is, to propose hypotheses
4. An ability to propose ways of testing hypotheses
5. An ability to apply principles to concrete situations
6. Accuracy of observation
7. Skill in use of the microscope and other essential tools
8. An ability to express effectively ideas relating to zoölogy[1]

There were already available techniques which are apparently valid for measuring the degree to which students have attained Objectives 1, 2, and 8. A test for microscope skill is described elsewhere in the BULLETIN, while ways of measuring 4, 5, and 6 will be presented at a later date. In this article will

[1] The formulation of these objectives and the work on the measurement has largely been done by W. M. Barrows, D. F. Miller, and J. W. Price of the Department of Zoölogy.

be described a method of measuring the ability of a student to draw inferences from the facts which have been presented.

By drawing inferences from facts is meant the ability to propose reasonable inferences which serve to interpret the related zoölogical facts obtained from observation, reading, or listening. It is essential in any such test that the student be presented with the necessary related facts which he is asked to interpret. Furthermore, in order that this may not be a measure of memory, it is necessary that these related facts be new to the student. If he previously has heard these facts interpreted or has read an interpretation, his response is likely to represent his memory of the interpretation made by someone else, which, of course, would not give a measure of his own ability to draw inferences from facts.

For an elementary class, the most productive source of new related zoölogical facts is the current research work in this field. A brief description of a recent experiment and its results presents material which is typical of the data which the student is expected to learn to interpret. Since the facts are new, successful performance is not a matter of mere memory. Hence, a series of such descriptions was prepared for use as test items. An illustration of one of the simpler items is:

> In autumn, when the pond is cold, frogs hibernate by burrowing into the bottom of the pond. In the spring a frog put into a tank of water goes to the bottom and attempts to burrow into the bottom of the tank.

After formulating a series of such items of varying difficulty, the next problem was the method of making the test objective. The ability to propose his own inferences without suggestion was desired for the student. It was hoped, however, that this ability might be highly correlated with the ability to select the best inference from a series of alternatives, which could easily be measured objectively by a multiple-response test. To learn whether the two abilities were highly correlated a group of sixty-six students was given two tests on the same day. The first test consisted of a series of items preceded by the following directions:

> Suppose in each of the exercises which follow that the facts given were observed or known to be true. Write in the space below the facts, the one inference which seems to you most reasonable in attempting to interpret the facts. The sample illustrates this procedure.

SAMPLE

Facts.—Starch when treated with iodine solution gives a blue color. When saliva is mixed with starch and left for a time the mixture no longer turns iodine solution blue.

Inference.—Saliva produces a change in starch.

Remember to write down the inference which seems most reasonable from the facts given.

These papers were then taken up and the students were given the same items in a multiple-response form. The following directions were used in this second form:

In each of the exercises below several facts are given. After the facts are suggested a number of inferences. Supposing these facts are true, select the inference which seems to you most reasonable and place a check mark in front of it as in the following sample.

SAMPLE

Starch when treated with iodine solution gives a blue color. When saliva is mixed with the starch and left for a time the mixture no longer turns iodine solution blue.

a. Saliva turns starch to sugar
b. Saliva turns iodine blue
√c. Saliva produces a change in starch
d. The color of saliva destroys the blue color
e. Saliva has no effect on grape sugar

Remember to check the inference which seems most reasonable from the facts given.

THE test items were the same in both tests, but the difference was in the ability required of the student. In the first test, he had to propose his own inferences; in the second, he had to select the most reasonable inference from a list of alternatives. A correlation between the results of these two tests shows whether the multiple-response test, which is much more easily scored, could be used in place of the first test as a measure of the ability to propose inferences. The method of scoring the first test will be described presently, showing that its score is highly reliable. The multiple-response test was scored in the usual manner. By the split-half method its reliability was found to be .90. However, when the results of the first test were correlated with the results of the second, the coefficient obtained was only .38. This indicates that there is some relationship between the tests, but by no means high enough for the multiple-response test to be used in place of the first test. The ability to select the most reasonable inference from a

given list is not the same as the ability to propose an original inference.

Since the objective multiple-response test could not be used to measure the ability to propose inferences, it became necessary to develop a reliable method of scoring the original inferences proposed by the students. To this end, several devices were tried. Each different inference proposed was typed upon a separate 3-inch by 5-inch card. No names were placed on the cards in order that the instructor might not be influenced. A set of these cards was given to an instructor in zoölogy who was asked to take each test item in turn and to sort the inferences into as many different levels of quality as he could identify, placing the best pile at the right and the poorest at the left. A record was made of the instructor's sorting, the cards were then shuffled and given to each of the other instructors in turn. The number of different levels of quality recorded varied from four in the case of one instructor to eleven in the case of another.

A given instructor's judgments were given numerical scores by assuming that all of the responses, if accurately measured on a scale of merit, would be normally distributed. By this means a value for each judgment can be obtained in terms of the standard deviation of this distribution, using the formula:

$$\frac{M_x}{\sigma_x} = \frac{Y_1 - Y_2}{F}$$

When

$\frac{M_x}{\sigma_x}$ = the mean value of a given judgment expressed in σ units,

Y_1 = the left ordinate of the normal curve under consideration,

Y_2 = the right ordinate,

and

F = the fractional part of all the cards which were placed in a given pile.

The values Y_1 and Y_2 may be quickly obtained from published tables of the normal curve.

After the scores had been calculated for each student on the basis of each instructor's judgment separately, the objectivity of the method was determined. It was found, if the ratings made by any three instructors chosen at random were used and each student's score obtained by averaging the scores given by the three instructors, that none of these scores differ more than $.3\sigma$

from the scores obtained by using the ratings of all the elementary instructors in the department. The ratings of three instructors, then, can be used to represent the combined judgment of the entire department as to the quality of the original inferences which have been proposed by the students.

ANOTHER device for evaluating these inferences, which is much simpler to use, was tried. Each instructor was given the set of cards and asked to place them in five piles according to merit, No. 5 representing the best inferences and No. 1 the poorest, attempting to place the cards so that the average quality of pile No. 5 was as much better than the average of No. 4 as No. 4 was better than No. 3, as No. 3 was better than No. 2, and as No. 2 was better than No. 1. This is the method of equal-appearing intervals. In scoring, an inference placed in pile No. 5 was given five points; one placed in No. 4 was given four points, and so on. The average scores obtained by this method were then correlated with the average scores obtained by the method previously described, giving a coefficient of .98. The simpler method of equal-appearing intervals, therefore, can be used in place of the σ method.

A final check on objectivity was made this Autumn Quarter. Several new instructors had been added to the department, none of whom had seen the set of inferences or had any idea as to their placement by the instructors last year. Three of these instructors were given the shuffled cards, and were asked to place them by the method of equal-appearing intervals. The average scores given by these three new instructors were correlated with the average scores given by the instructors the year before. The resulting correlation coefficient of .96 indicates that the scoring by three zoölogy instructors is not likely to vary from year to year nor to depend upon the particular personnel of the staff, and that the merit of an inference is not the purely subjective quality often claimed.

In addition to making certain of the objectivity of the scoring, it is necessary to know whether the number of test items is adequate for the determination of the student's ability or the interpretation of zoölogical facts. Since the items used represent only a sample of all the facts which he might conceivably be expected to interpret, the reliability of the test depends upon the adequacy of this sample, and was determined by the split-half

method and the use of the Spearman-Brown formula. The reliability coefficient thus obtained was .95. It thus appears that the test is both reliable and objective.

Several persons who have constructed tests have maintained that all of the important objectives were so closely correlated with information that an information test was all that was necessary in measuring achievement. This is not true in zoölogy. The correlation between a test covering the important zoölogical information and this test of proposing inferences was found to be only .29, entirely too low to justify the use of the information test alone as a measure of both objectives. If the ability to propose inferences is an important objective of zoölogy, we must have a measure for it, for we cannot depend upon information tests to indicate the student's attainment of this objective.

A final problem is now under investigation. It is very necessary that this method of measurement does not become a test of memory, hence the items must always be new to the students. This is only possible by selecting new items from current research for each new test. If new items are continually used, there is a danger that they will be of differing difficulty from those used previously, resulting in the condition that no evidence of student growth in this function will be obtained. We are now trying to formulate a method by which the difficulty of a new item can be predicted accurately in advance in order that the test may consist of new items while at the same time it can be used to measure student progress.

* * *

Test Method Helps, Number Two, has recently been published by the World Book Company. This booklet, by Ross O. Runnels, is a "manual for determining the equivalence of mental ages obtained from group intelligence tests." The equivalent mental-age norms have been worked out for National Intelligence Tests: Scale A, Form 1; Haggerty Intelligence Examination, Delta 2; Otis Group Intelligence Scale, Advanced, Form A; Terman Group Test of Mental Ability, Form A; and Otis Self-administering Test, Higher, Form A. The equivalent norms will be useful in comparing the results from intelligence tests since significant differences have long been noted in the mental ages obtained from various tests.

Objective Measurements in English

By ELVA ANNE LYON

OBJECTIVE measurements in English have been used, during the Autumn Quarter in the College of Agriculture at Ohio State University, as one of the factors in the student's analysis of his abilities and aptitudes. The experiment is a further development of the policy already in operation for a year in several departments of that college whereby objective measurements are used with the aim of providing for and encouraging the fullest possible development of each individual. This necessitates taking into account all available factors of aptitude and ability and of making available to the student the results of the objective measuring of these factors so that he may share with the faculty the responsibility of using these known facts in developing a successful college career, which leads, in turn, to satisfactory participation in a wisely chosen profession.

This program of self-discovery for the Freshmen forms the core of the one course required of all of them, the Survey of Agriculture conducted by Mr. Nisonger, the junior dean. It was at his invitation that the English experimentation has been carried out in this course. Much of the credit for whatever value the procedure has had must be given to his interest, suggestions, and co-operation.

In the College of Agriculture, the regularly required English course does not begin until the second quarter. The students, however, were given no chance to fall into the habit of thinking that until the regular English course started English was of no importance. In the survey course and elsewhere they were impressed with the fact that expressing their thought correctly and accurately was essential to successful college work. Certain of the objective pre-tests in botany and zoölogy required the interpretation of observed facts and the drawing of inferences from given situations. Difficulty in expressing ideas clearly was shown to be a definite handicap on such tests. Furthermore, the general attitude of instructors that botany, for instance, is not good botany unless it is well expressed gives the students unmistakable evidence that English is a practical subject that functions in all their college work and not a departmentalized routine that operates only during the English class.

The felt need for improvement of English thus stimulated set the stage for the introduction of the analytical work early in the year. But before describing the program as it was carried out, it will be well to define a little more clearly the objectives of the experiment, the data, and the methods.

The aim was not limited to ascertaining what errors in English the present group of students made at the beginning and to measuring how much improvement could be effected in a given period. It was felt that greater service could be rendered by locating at the same time, and as definitely as possible, the factors which have limited the students' facility in correct use of English, to isolate as far as possible the principles involved in helping the student to locate his English difficulties and to work out a satisfactory solution for them, and to seek for long-term improvements and results rather than for quick, and probably temporary, panaceas. The data collected and the methods used have been based on an analysis of the problems involved in the situation and in the objectives, both immediate and remote, that have been tentatively set for the study. The present purpose is to describe, rather than to evaluate, the use of objective methods in defining the problems, suggesting solutions, and how it is intended to use them in measuring results.

The data used thus far in the study have been the following:

1. The student's high-school record, including his standing in his class, that is, whether in the upper, middle, or lower third; the grades received in his high-school English courses; and his attitude toward his high-school English
2. The student's rating on tests taken upon his entrance to the University: a psychological test, and a reading test
3. The Pressey Diagnostic English Test
4. A written paper prepared for the course, Survey of Agriculture
5. The student's statement as to what he feels to be his chief writing difficulties

The data of the student's high-school record were obtained from the forms filled out in making application for admission to the University. His attitude toward English was found, if found at all, in his response to questions on the admission blank:

> What studies did you like best? Give reasons.
> What studies did you spend the most time on?
> What studies did you dislike most, or like the least? Give reasons.
> What studies did you find most difficult?
> What studies did you work hardest on?

This, of course, gave the student's reaction to English only when it occurred in one of these groups, but it is surprising to note, from the results given later, how frequently English appeared in the answer to the third or fourth of these questions. The data on the psychological and reading tests were obtained from the report of these tests furnished the College of Agriculture by the Department of Psychology. Only in the case of the Pressey English tests and the student's statement of his "felt writing difficulties" were data used that were not already available from routine, administrative sources, or as a part of regular class assignments in the survey course.

THE first project upon which the student worked in the survey course was keeping a record of his expenditure of time during his first week of college work. A form was furnished which provided spaces for recording the hours he spent each day in the classroom or laboratory, in the preparation of lessons, in outside activities (which he was to list), in outside reading, in outside work, and in commuting. A day or two before a report of this project was to be handed in, the student was given this assignment sheet to guide his writing:

SUGGESTIONS FOR THE DISCUSSION OF YOUR TIME BUDGET

In the activity record of your first week in college you have the basis for building efficient working methods which will determine to a great extent your success as a college student. A thorough and critical examination of this time budget should enable you to detect which activities have been allotted sufficient time and which usurp time that should be given to things more worth while to you. Test your program by the questions suggested for thought which accompanied your record sheet, and by others which occur to you.

One of the best ways of thinking a problem through clearly and thoroughly is to write it out so that it will convey your ideas to your reader in a direct and forceful way. You are asked to do this. Make it clear to him why you are organizing your time the way you are, what there is in the record of your first week that you wish to continue, and what you will attempt to adjust to meet your needs better.

Your written discussion will be evaluated on the following points:

1. The thoroughness of your analysis of your time budget as it relates to your problems.
2. The organization of your discussion of this analysis.
3. Your choice of language.
4. The conciseness and clearness of your expression.

After this written assignment was handed in, a class period was used for administering the Pressey English Test. A mimeographed report of the results of the test was provided for the

students, and the tests themselves were returned and discussed, with stress on the reasons for the correct forms or the specific grammatical rules involved, so that this information would be most likely to function as a working tool in the student's writing. The report contained a statement of the mean score made by the group of 164 on each of the four parts of the test, so that each student could compare his rating with the average for the group. The mean score of the group for the combined scores of the four parts of the test was 69.3. The mean score made last May by about nine hundred high-school Seniors in twelve high schools in Ohio was 80.[1] The agriculture students were told that, in consideration of the facilities that were being provided to assist them with their English difficulties, it seemed reasonable to expect them to bring their average up to 80 by the end of the quarter. The report also contained a difficulty analysis of the items on each part of the test, compiled by tabulating the frequency with which each had been missed and then arranging the items in order from the one most frequently missed to the one of lowest frequency. Also, a grouping of errors according to type was made, and methods of correction indicated. The value of this procedure was to impress the student that his English was not hopeless, but that attention to a few major types of error would eliminate most of his trouble. A résumé of this information is given in the paragraphs which follow for those teachers interested in diagnosing the errors of their students by a similar method, or who wish to use the findings of this study as suggestive of the principles they should stress in their own classes or to indicate the errors likely to be found most prevalent in the writing of their students.

SENTENCE STRUCTURE—

Over 40 per cent of the total errors consisted in failures to recognize incomplete sentences, or the punctuation of two or more complete sentences as if they were one sentence.

Thirty-five per cent of the errors involved incorrect or indefinite antecedents of pronouns.

Twenty-two per cent of the errors consisted of incorrect placing of words, phrases, or clauses with relation to the word they modified.

GRAMMAR—

Fourteen per cent of the errors in grammar consisted of violations of agreement between the subject and predicate in person and number, chiefly following *there is*, when subject nouns are joined by *and*, *or*, or *nor*, or when

[1] Tyler, R. W. "High-School Pupils of Today," *Educational Research Bulletin*, IX (October 22, 1930), pp. 409-11.

the verb is made to agree with a noun standing immediately before it which is not the subject.

Sixteen per cent of the errors involved the correct number of pronouns and verbs following such words as *each, neither, everybody, none, a person.*

Eighteen per cent of the errors consisted in confusion of the nominative and objective cases of pronouns in dependent clauses, and as objects of verbs or prepositions in such usage as *to us girls* and *between Mary and me.*

Incorrect adverbial form or usage constituted 12 per cent of the errors.

CAPITALIZATION—

Almost 10 per cent of the errors in capitalization occurred in this sentence, "He lives in the West, a few miles north of our old home"; and 18 per cent involved these two sentences, "I will write to Major Cox and Professor Bell," and "He had been a professor, but enlisted and rose to the rank of major."

Other items showing high frequency of error were the capitalization of such words as *high school, company,* and *club,* adjectives or other nouns derived from proper nouns, capitalization in titles of books and in direct quotations, both continuous and broken.

PUNCTUATION—

The highest frequency of error involved these usages: The apostrophe to indicate omitted letters and to show possession; the setting off of parenthetical expressions, and conjunctive adverbs such as *however* and *nevertheless;* the punctuation to be used with "as follows" in introducing an enumeration; punctuation involved in distinguishing between restrictive and non-restrictive clauses; and distinguishing between the use of the comma and the semicolon between the members of a compound sentence.

The written assignment was carefully read, the errors indicated, and a rating given according to the following standards:

0.0 to 1.0—

Extremely poor work, even for beginning; the subject incompletely treated and badly organized; spelling errors prevalent; incomplete or run-on sentences; errors in tense or principal parts of verbs, or in agreement with subject in person and number; shift in pronouns, or incorrect antecedent and verb relationships with the pronoun.

1.1 to 2.0—

A fair beginning; subject-matter reasonably complete and organization satisfactory for the most part; fairly good command of sentence structure; some error in agreement, in pronouns and adverbs.

2.1 to 3.0—

A good beginning; treatment of subject-matter adequate and organization good; not more than two gross errors in fundamentals; sentences must in general be well constructed and show some variety.

3.1 to 4.0—

Excellent work for beginning; free from all gross errors; treatment of subject shows some originality; choice of words good; sentences interesting, with variety in structure, and punctuated for clearness of meaning and ease in reading. The paper presents a neat, well-planned, workmanlike appearance.

The limited class time permitted almost no discussion of the written assignments after they had been returned, but a weekly conference period was announced to which all students were welcome to come to talk over with the writer the specific difficulties which the Pressey test and the writing assignment had pointed out. The students who had made a low rating (one which fell approximately in the lower quartile on either or both) were especially urged to take advantage of this opportunity, but in no case was the conference made compulsory. A little more than a half-day is required each week to take care of those who take advantage of this offer to talk over their difficulties and to develop methods of overcoming them. The system has not been in operation long enough to enable the drawing of conclusions as to the steadfastness of this interest on the part of the student in improving his English, or to test definitely the results of the method. But quite apart from results, it seems significant that so many students have voluntarily shouldered the responsibility for improvement in their English when once their weaknesses have been located and the way provided for self-help.

At the present time a second report is to be issued to the students in which is incorporated the results of the student's statement of his "felt difficulties" in writing together with the investigator's analysis of his difficulties as his writing shows them, and suggested general methods for overcoming these troublesome factors. Before the students had read the critical comments on their returned papers, they were asked to list the four or five difficulties which bothered them most when they were writing, whether or not these difficulties showed up as errors in the finished product. The students' "felt difficulties" are, in order of importance according to frequency of mention: punctuation, sentence structure, organizing material and paragraphing, spelling, clearness and conciseness of expression, limited vocabulary, the grammar involved in the use of pronouns, verbs, and adverbs, and legibility and good form in the manuscript. The analysis of the students' writing was completed before the statements of felt difficulty were obtained; for that reason, as well as from an attempt to analyze the category of errors to agree more or less closely with the Pressey test, the categories on the two tabulations do not coincide sufficiently to

be strictly comparable. This analysis gave frequency of error in this order: sentence structure, including conciseness of expression; grammar, adverbs, pronouns, verbs, prepositions; punctuation; and spelling.

The class or conference discussions of some of these difficulties as the test located them or as the first report to the students treated them are, tentatively considered, sufficient. The second report will be confined to suggested self-help on the topics of spelling; legibility and form of the manuscript; vocabulary building; development of clearness and conciseness of expression; some further grammatical work on pronouns, verbs, and adverbs; and the organizing of material and development of paragraphs. Shortly after this report is in the hands of the students another writing assignment for the survey course will be read and analyzed to note the improvement made and the errors which still persist, and which will need further treatment. At the end of the quarter objective tests will again be administered, comparable to the ones used at the beginning, to indicate definitely what changes have taken place in the student's English equipment. All data accruing from this experimental work will be available to the regular English instructors the following quarter to use as they see fit.

So much for the program of objective measurement of English as a factor in the student's self-analysis and self-help with regard to the abilities which are closely related to college success. Further treatment of the data available will give some idea of what the study seems, at present, to reveal in the broader sense concerning the English preparation of these college Freshmen.

The use of an objective test in English which consists of testing the student's ability to detect error in grammar and sentence structure on a printed form, or to punctuate and capitalize correctly sentences already written, always raises the question as to the relation of success or failure on such tests to the student's ability to apply this information in his own writing. The average length of the composition read in this study was about two hundred words, certainly much too short to give a high degree of expectancy of finding every error made on the Pressey test duplicated on the written work. The analysis of many thousand words of the student's written work

would certainly be needed for a comparative study of that type. The present investigation did make an attempt to compare the general quality of the work of the student on the test and on his written work by correlating his test score with the rating given his composition on the four-point scale already described which does, of course, involve much more of the subjective element than the test score. The summary which follows indicates the correlation coefficient found between these two ratings, and between various other ratings that were thought to be of sufficient importance to warrant this statistical treatment:

Ratings Correlated	Correlation Coefficient
Total score on Pressey test correlated with	
Composition rating	.7
Intelligence percentile	.6
Reading percentile	.7
Mean of high-school English marks	.2
Composition rating and mean of high-school English marks	.1
Intercorrelation on the Pressey test	
Capitalization and punctuation	.4
Sentence structure and grammar	.6
Sentence structure and punctuation	.6

The writer is suspicious of the relatively high correlation obtained between the Pressey score and the composition rating, and until much more experimental evidence is at hand would much prefer attributing it to "luck" or "chance" rather than claiming that a highly efficient method of rating writing objectively has been discovered. It may be noted that there is a decidedly higher correlation between the ratings established by wholly or partially objective measurements used at the University than between the grades received for high-school work, which may or may not have been arrived at by objective methods, and the ratings received on university tests. Only a careful study of the student's subsequent university attainment can indicate whether the high-school or the university rating is most accurately prognostic of his success at the University.

Correlation coefficients between the parts of the Pressey test that are only moderately high might be taken to indicate that English ability is not general and uniform, but rather is made up of specific abilities. A student, for instance, may be reasonably proficient in capitalization and punctuation and at the same time be quite lacking in the ability to write correct sentences or to use the parts of speech with due regard to grammatical rules. Developing integrated writing ability would

seem to indicate the necessity of locating the specific difficulties and attacking them first.

The rather low mean score which this group of Freshmen made on the Pressey test raised the question as to whether the group was generally low or whether English was the exception. The mean ranking of the group on the Intelligence Percentile Scale and the Reading Percentile Scale was found to be 45.9 and 42.7, respectively. The ranking of these students in their high-school classes was found to be as follows: 36 per cent were in the upper third of their class; 52 per cent, in the middle third; and 13 per cent, in the lower third. These figures indicate that low English rating probably cannot be explained entirely on the basis of low general scholarship.

The attitude toward high-school English was next considered. On the admission blanks, English was listed among the studies liked best or studied most by twenty-four students, while 116 listed English as their most disliked or most difficult subject. The reasons for the dislike of English most frequently given were that it was not practical, or did not have any connection with other studies, that they could not understand it, that they had difficulty in writing, or that they could not learn strings of dates or memorize poems. In other words, for this group of students, coming from more than a hundred high schools, English "did not get across." The present experiment seems to indicate, even this early, that it is possible to detect by objective means some of the factors involved in this failure, and it hopes to formulate before it is concluded some methods for overcoming these difficulties and to test their value by objective measurements.

* * *

That interest in the testing movement has become marked in colleges and universities is shown by a recent grant of $500,000 made by the General Education Board to the American Council on Education. This subvention, which is to be expended over a ten-year period, will be used to develop achievement tests which will be employed for the rating of students in American colleges and universities.

What High-School Pupils Forget

By RALPH W. TYLER

THE statement is frequently made that what little high-school students learn is quickly forgotten. In some cases, this claim has been supported by test records showing that pupils made distinctly lower scores a year after taking a given course than at the close of the course. In considering the nature of this forgetting, it seemed probable to the writer that information gained in a course which is seldom needed thereafter would, in general, be forgotten most quickly. On the other hand, explanations which can be commonly applied to the everyday happenings of life and methods of thinking which are frequently used might be expected to be retained by high-school pupils with much less loss.

To investigate this hypothesis an experiment was planned and conducted through the co-operation of the teachers in two high schools in Columbus County, North Carolina. The ninth grade in each school was composed of pupils who had taken general science the previous year. Appropriate tests given in September would indicate what was retained by these pupils after four months had elapsed since completing the general-science course. Since this subject was not taught in the ninth grade, similar tests given eight months later would show the amount of forgetting during a school year in which the pupils were not taking science.

To measure the amount of information which the pupils remember from their general science, the Ruch-Popenoe General Science Test was used. Form A of this test was given in September and repeated eight months later. To measure the ability to explain everyday scientific phenomena and to generalize from given scientific data, two new tests were developed. The explanation test was a multiple-choice test consisting of fifty items. Each item consisted of a statement of a commonly observed scientific phenomenon followed by five alternative explanations, the best one of which was to be checked. An illustration of one of the test items is:

It is less likely to frost on cloudy nights in late autumn than on clear nights, because:

a. It will not frost when the air is damp.

b. The clouds reduce the radiation of heat from the earth.

c. Clouds reduce the amount of heat received from the sun.
d. It is always colder on clear nights.
e. There is more moonlight on clear nights than on cloudy nights.

This test was given at the same time that the Ruch-Popenoe test was given and was also repeated eight months later.

The third test was one requiring generalization from given facts. Each item consisted of a brief description of an experiment. The results obtained from the experiment were given, followed by five alternative inferences which might be drawn from these results. The pupils were asked to check the inference which was most reasonable from the facts given and the one least likely to be true. An example of one of the test items is:

A piece of glass immersed in water comes out wet, when immersed in mercury it comes out without any mercury adhering to it. When a small glass tube is placed in a basin of water, the water rises in the tube above the level of the water in the basin. When the tube is placed in a dish of mercury, the level of the mercury in the tube is not as high as the mercury in the dish.

a. The attraction between particles of mercury and glass is greater than the attraction between mercury particles, hence the mercury is pulled down in the tube. With water and glass the reverse is true.

b. The attraction between particles of water and glass is greater than the attraction between water particles, hence the water is pulled up so that there is a greater contact with the glass. With mercury the reverse is true.

c. This is an example of osmosis which causes the water to rise and the mercury to be depressed. It only takes place in small tubes.

d. The surface tension of water is much greater than that of mercury as shown by the fact that water wets glass. Hence the surface of the water rises in the tube. With mercury the reverse is true.

e. The molecules of water must be much smaller than those of mercury so that they can rise in small tubes easily. The large mercury molecules encounter difficulty in moving in the tube.

This test was also given in September and repeated eight months later. The reliability of the two new tests were .89 and .91, respectively. Table I gives the mean scores made on the initial tests as well as those received on the final tests.

It is probably to be expected that there was a distinct drop between the initial and final scores on the Ruch-Popenoe test which measures largely information. The mean score in September was 27.9 and was only 16.8 in May. On the other hand, there was little difference between the initial and the final mean scores on the other two tests. It seems apparent that for these pupils the greatest loss during the year was in

information about science, whereas there was little or no loss in the ability to explain everyday phenomena and the ability to generalize from given facts.

If corresponding results be found quite generally, they are of great significance. There would then be justification for concluding that the permanent values of high-school science are not so much the information retained as the ability to explain in scientific terms and the ability to generalize from

TABLE I

AVERAGE SCORES ON INITIAL AND FINAL TESTS

Test	Number of Pupils	Mean Initial Score	Mean Final Score	Mean Loss in Eight Months	Probable Error of Mean Loss
(1)	(2)	(3)	(4)	(5)	(6)
Ruch-Popenoe	68	27.9	16.8	11.1	1.7
Explanations	68	23.4	21.6	1.8	1.1
Generalizations	68	21.7	20.5	1.2	.9

known facts. If this be true, the common method of determining what pupils obtain from a high-school education by giving them a range-of-information test is incomplete. These other abilities less likely to fade overnight must also be plumbed before we are justified in concluding that high-school students quickly lose all that they have learned in their four years of training. They may forget information which is little used thereafter, but it is quite possible that they retain those methods of thinking in a given field for which there is more frequent occasion for use.

* * *

The Supervisory Service Bureau of the School of Education of the University of Wisconsin at Madison has recently published the Torgerson Diagnostic Teacher Rating Scale which consists of a series of statements describing rather definitely ninety practices followed by teachers. The supervisor checks those statements which she believes apply to the teacher being rated. The validity of the scale has not been determined although the author states that "every aspect of teaching listed in the scale correlates positively with teaching success." The scale is reported to be highly objective and reliable.

A Test of Skill in Using a Microscope

By RALPH W. TYLER

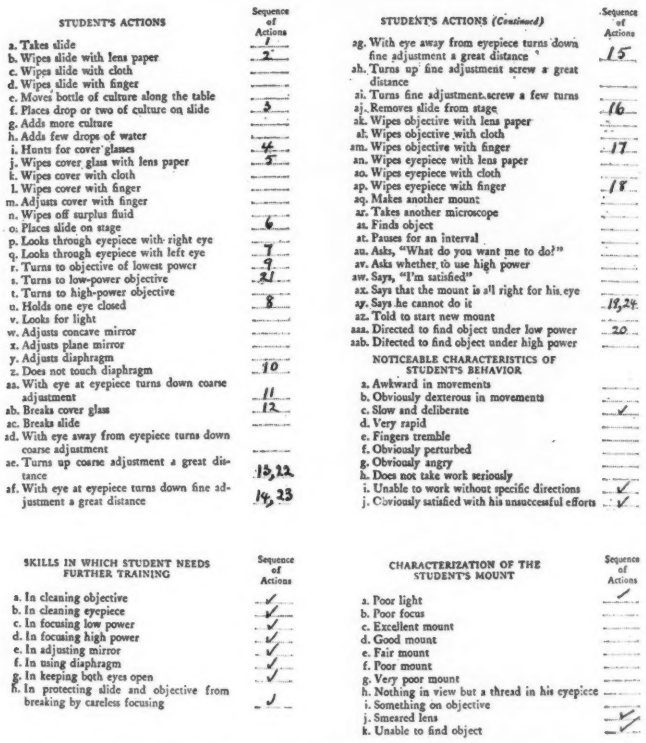
ADEQUATE and reliable methods for testing the amount of information which a student remembers have been developed recently in many fields of subject-matter. Equivalent progress in developing other types of measurements has not been made. Few tests are yet available for measuring the degree of laboratory skill attained by the students, yet in the natural sciences one objective is to develop certain laboratory skills which are essential for getting first-hand contact with scientific phenomena.

The necessity for a test in microscope skill in botany and zoölogy classes at Ohio State University became apparent when it was found that some students in advanced classes were unable to use a microscope effectively. These students had taken two courses in which they had been expected to use the microscope, yet a few of them were totally unable to manipulate the instrument. They had been unable in these two courses to make any first-hand observations of the minute structures in which most of the life processes of plants and animals take place. To discover these students who are not learning to use the microscope early, to determine the nature of their difficulties, and to provide the necessary remedial instruction so that they may gain in the essential proficiency it was necessary to devise a test of skill in using the microscope.[1] Such a test should be administered after the students had been given the usual class training in using a microscope, in order to identify those who had not attained the desired skill.

The writer made a series of observations of the reactions of individual students for the purpose of determining the types of difficulties students encounter in using the microscope. Thirty-three students in Botany 402, the second half of the elementary course, were selected for observation. Using a room which was unoccupied for the day, two microscopes were placed on a table together with yeast culture, slides, covers, cloth, and lens paper. The writer placed himself in a position which gave him a good view of the actions of anyone at the table. The instructor, Miss Pearle Williams, then called in from the laboratory one

[1] The construction of the test was suggested by H. C. Sampson of the Department of Botany and was developed through his help and the co-operation of W. M. Barrows of the Department of Zoölogy.

student at a time. When the student came in she told him that the bottle contained some yeast culture and asked him, as quickly as he could, to find a yeast cell under the microscope and to show the cell to her when he had found it.



STUDENT'S ACTIONS	Sequence of Actions
a. Takes slide	1
b. Wipes slide with lens paper	2
c. Wipes slide with cloth	
d. Wipes slide with finger	
e. Moves bottle of culture along the table	
f. Places drop or two of culture on slide	3
g. Adds more culture	
h. Adds few drops of water	
i. Hunts for cover glasses	4
j. Wipes cover glass with lens paper	5
k. Wipes cover with cloth	
l. Wipes cover with finger	
m. Adjusts cover with finger	
n. Wipes off surplus fluid	
o. Places slide on stage	6
p. Looks through eyepiece with right eye	
q. Looks through eyepiece with left eye	7
r. Turns to objective of lowest power	9
s. Turns to low-power objective	21
t. Turns to high-power objective	
u. Holds one eye closed	8
v. Looks for light	
w. Adjusts concave mirror	
x. Adjusts plane mirror	
y. Adjusts diaphragm	
z. Does not touch diaphragm	10
aa. With eye at eyepiece turns down coarse adjustment	11
ab. Breaks cover glass	12
ac. Breaks slide	
ad. With eye away from eyepiece turns down coarse adjustment	
ae. Turns up coarse adjustment a great distance	13,22
af. With eye at eyepiece turns down fine adjustment a great distance	14, 23

STUDENT'S ACTIONS (Continued)	Sequence of Actions
ag. With eye away from eyepiece turns down fine adjustment a great distance	15
ah. Turns up fine adjustment screw a great distance	
ai. Turns fine adjustment screw a few turns	
aj. Removes slide from stage	16
ak. Wipes objective with lens paper	
al. Wipes objective with cloth	
am. Wipes objective with finger	17
an. Wipes eyepiece with lens paper	
ao. Wipes eyepiece with cloth	
ap. Wipes eyepiece with finger	18
aq. Makes another mount	
ar. Takes another microscope	
as. Finds object	
at. Pauses for an interval	
au. Asks, "What do you want me to do?"	
av. Asks whether to use high power	
aw. Says, "I'm satisfied"	
ax. Says that the mount is all right for his eye	
ay. Says he cannot do it	19,24
az. Told to start new mount	
aaa. Directed to find object under low power	20
aab. Directed to find object under high power	

NOTICEABLE CHARACTERISTICS OF STUDENT'S BEHAVIOR	
a. Awkward in movements	
b. Obviously dexterous in movements	
c. Slow and deliberate	✓
d. Very rapid	
e. Fingers tremble	
f. Obviously perturbed	
g. Obviously angry	
h. Does not take work seriously	
i. Unable to work without specific directions	✓
j. Obviously satisfied with his unsuccessful efforts	✓

SKILLS IN WHICH STUDENT NEEDS FURTHER TRAINING	Sequence of Actions
a. In cleaning objective	✓
b. In cleaning eyepiece	✓
c. In focusing low power	✓
d. In focusing high power	✓
e. In adjusting mirror	✓
f. In using diaphragm	✓
g. In keeping both eyes open	✓
h. In protecting slide and objective from breaking by careless focusing	✓

CHARACTERIZATION OF THE STUDENT'S MOUNT	Sequence of Actions
a. Poor light	✓
b. Poor focus	
c. Excellent mount	
d. Good mount	
e. Fair mount	
f. Poor mount	
g. Very poor mount	
h. Nothing in view but a thread in his eyepiece	
i. Something on objective	
j. Smeared lens	✓
k. Unable to find object	✓

FIG. 1. Check List of Student Reactions in Finding an Object under the Microscope

As each student began, the writer noted the time and made a record of the sequence of actions. When the student finished, the time was recorded together with a brief characterization of the quality of the student's mount. From these thirty-three observations a preliminary list of actions, both desirable and undesirable, was obtained. This preliminary list was supplemented by more than fifty additional observations. A portion of the final form of the check list is shown in Figure 1.

THE check list as now used requires little writing. The arrangement is substantially the same as when it was first developed. A botany student is asked to find a yeast cell, whereas a zoölogy student is asked to find a blood cell. A record of the time is kept, but instead of writing down his actions the check list contains all of the actions both desirable and undesirable which have been observed thus far. The observer records the sequence of actions by placing the figure *1* after the description of the student's first action, a figure *2* after his second, and so on. The sample reproduced here has been filled out for illustration. By reading the actions in the order they are numbered one gets a detailed description of the student's procedure in finding an object under the microscope.

An examination of the sample will serve to indicate the usefulness of the test. It is clear that this student is very deficient in microscope skill. He used eleven minutes without finding the cell, whereas the better students find an object in approximately two minutes. He does not adjust mirror nor diaphragm, he closes one eye and turns down the coarse adjustment, while his eye is at eyepiece—all habits which must be remedied before he can use a microscope skillfully. As would be expected, he has poor light on the object, has a smeared objective, and is unable to find the object. He needs training in each of the specific phases of microscope skill. Hence, this test record gives the instructor an analysis of the student's behavior which permits the selection of the individual students who need training and a determination of the specific phases of training which each student needs.

The most promising part of the procedure is the discovery that students deficient in microscope skill can be quickly trained to reasonable proficiency. Miss Blanche B. Montgomery has had charge of the remedial training in the Department of Zoölogy. She has been giving the test followed by remedial instruction to a large number of students. She finds that ten minutes of her time on the first day for instruction, followed by five minutes for a check a week later, will provide the necessary guidance for more than half of the students. After the ten minutes of instruction the student spends several periods in practice, with a result that more than 90 per cent of the students have become sufficiently skillful to use the microscope effectively.

This microscope test is, of course, an individual test which requires a certain amount of the instructor's time. Ordinarily, ten to twelve students can be tested in an hour, but to use such an individual test with each of the seven hundred students in Botany 401 would require nearly two weeks for one instructor. Only the persons who are having difficulty in using the microscope need to be observed. By means of an easily administered group test those who should take the individual test may be selected. In the group test, the entire section is in the laboratory seated at microscope tables each of which is provided with the necessary materials. The instructor informs the class that they are to see how rapidly they can find an object under the microscope. They are all to commence when he gives the signal and to stop when he gives the second signal.

The second signal is given at the end of three minutes, and the instructor quickly passes from microscope to microscope noting those students who are unable to find the object in three minutes. This same group test is given a second time several days later. Those students who were not able to find an object in three minutes in both tests are then given the individual test. This group test requires only ten or fifteen minutes to administer, but by means of it the instructor identifies about three-fourths of the class who are already proficient with the microscope. This means that in a class of forty only ten need to be given the individual test. These ten tests can be given in an hour, which is not a heavy burden on any instructor.

The microscope test is not a paper-and-pencil test, but it is an objective device for determining the degree of skill possessed by students in biology. By its administration a record is obtained which permits the instructor to determine the type of remedial training needed and to give that training with a minimum expenditure of time and effort. The student thus trained has acquired skill in the use of an essential tool of biology.

EDITORIAL COMMENT

To MANY students of education "objective measurement" is synonymous with a paper-and-pencil test, such as a list of true-false statements. Where used with such a restricted connotation objective measurement is not of paramount importance.

No such restriction should be placed upon the term. The members of the Achievement-Test Division of the Bureau of Educational Research in the development of objective tests are concerned with techniques for obtaining any objective evidence of the degree to which students are attaining the important goals of education. If one purpose of a course in literature is the development of a desire to read good books, then we need a method for finding out the degree to which students are developing this desire. If a purpose of hygiene is the development of healthful personal habits, we need techniques for determining how adequately such habits are being developed among school children. By no other means can we adequately evaluate our teaching procedures, can we diagnose our pupils' difficulties, and can we determine the effectiveness of remedial training.

Objective measurement so conceived has two major tasks, the invention of ways of measuring all important aspects of educational development and the elimination of unreliable, widely varying, subjective estimates of a student's attainment. To invent ways of measuring all important aspects of educational development means that we must not restrict ourselves to paper-and-pencil tests alone—least of all to some particular form of paper-and-pencil test. Valid evidence of a student's attainment is of first importance whether that demands observation of the student's habits, evaluation of his oral speech, or what not. After we have invented ways of getting valid evidence, the attempt to eliminate subjectivity and to reduce the cost and difficulty of measurement without reducing the validity of the measures should be made. These are the research tasks of those engaged in objective measurement.

This broader definition of objective measurement makes clear its fundamental importance. The ultimate value of any method of teaching, or scheme of administration, or of any educational device can only be determined in terms of measures of the degree to which this method or this device facilitates the attainment of the desired objectives of education. R. W. T.

READINGS

"Student Help in the Growth of the Teacher," *Educational Method*, X (October, 1930), pp. 30-33.

This teacher sought a means of evaluating herself as a teacher. She could trust neither her own judgment nor the professional promotion which she had received. She wondered how she had measured up to the standards of pupil improvement. She asked her students to criticize her as a teacher and person, the course of study, and the method of conducting it. She urged them to be truthful and to state the case favorably or adversely just as they thought. They were told not to sign their names as she did not want to know who wrote the papers.

The papers were analyzed for a list of qualities. Twenty-one qualities were recognized and approved by the students. These qualities with terms by which the students described the qualities are given in the article. She also presents a list of adverse criticisms.

BOARDMAN, CHARLES W. "An Analysis of Pupil Ratings of High School Teachers," *Educational Administration and Supervision*, XVI (September, 1930), pp. 440-46.

The writer of this study obtained and analyzed some of the factors which may enter into pupils' judgments of teachers and compared pupils' judgments with those of supervisors and teachers. The pupils were given sheets of paper and asked to rank their teachers during the current semester on the following questions: (1) "For which teacher do you work the hardest?" (2) "Which teacher do you like the most?" (3) "Which teacher has the best order or discipline?" (4) "From which teacher do you learn the most?" (5) "Which teacher do you think is the best or most efficient in teaching?" The supervisors also ranked the teachers for their teaching efficiency. Furthermore, each teacher ranked as many of his associates as he felt competent to appraise in terms of his own training and experience.

The statistical technique of correlation was used to determine the relationship among the rankings of the various groups and the rankings of the pupils on the five questions.

The coefficients seem to indicate the pupils' judgments of teaching efficiency are based upon factors which supervisors would recognize as important. Though the rankings do not correlate to a high degree with supervisors' rankings, there is no evidence that the latter are a valid criterion for judging the pupils' rankings. It may be that pupils possess knowledge of their teachers' efficiency which would be valuable to supervisors.

CURETON, EDWARD E., AND DUNLAP, JACK W. "Scoring the Rearrangement or Continuity Test," *School Review*, XXXVIII (October, 1930), pp. 613-16.

The continuity test is a special adaptation of the rearrangement test as applied to the chronological order of historic events. The writers of this article believe that the best measure of the pupil's knowledge of this type of material is the correlation between his own ranking and the correct ranking. Spearman's rank-difference correlation formula is used in a modified form

to obtain percentage results. Since such computation for each score would be laborious, they have worked out a nomograph whereby the percentage score can be determined by connecting two points and extending the line to the percentage scale. The two points connected are the sum of the differences between the correct order and the pupil's order squared, and the number of items on the test.

RULON, PHILLIP JUSTIN, AND ARDEN, WESLEY. "A Scoring Technique for Tests Having Multiple Item-Weightings," *Personnel Journal*, IX (October, 1930), pp. 235-41.

In such a test as the Strong Vocational Interest Blank, in which there are 420 possible responses evaluated in twenty different ways for twenty different vocations, the work of scoring and tabulating becomes prodigious and expensive. The Hollerith Printing Tabulator is a machine which alleviates this burdensome work. This article gives a helpful description of how the Hollerith machine may be used to score weighted responses and tabulate the results of such complicated tests.

CHAUNCEY, M. R. "Must All Measures of School Work Be Quantitative?" *Journal of Educational Research*, XXII (October, 1930), pp. 205-10.

Eight techniques are listed for measuring the degree of success of pupils in performing activities, observation, personal interview, group conference, diary, question blank, check list, rating scale, and standardized objective tests. Since many of the results of education cannot at present be quantitatively measured, this article makes an appeal for the inclusion of other than purely quantitative techniques in courses in measurement for teachers in training.

CAHOON, G. P. "A Tabulation and Analysis Sheet for New Type Tests," *University High School Journal*, X (August, 1930), pp. 160-75.

A cross-hatch tabulation technique for recording the responses of individuals to the items of a test is presented. It places the entire results of a class test on one sheet of paper, which enables one to analyze easily the items of the test and the pupils' responses. It may be used to show relative pupil achievement and to diagnose pupil and class difficulties.

In Paper Covers

PRESSON, JOHN M. *Achievement Tests in Biology for Secondary School Use Based upon an Analysis of the Content of the Subject.* A Doctor's dissertation on file in the Library of the University of Pennsylvania, 1930. 149 pp.

The purpose of this study is the construction of standardized achievement tests in the field of secondary biology. The validity of the content of the tests was established by a careful analysis of the subject-matter that is taught in the tenth grade of high schools. Separate tests have been devised for plant biology, and animal and human biology. The reliability of two forms of each test is fairly high. Administering both parts of each test requires 32 minutes, which means the test can be given and scored in one class period.

As an information test, its validity is well established. However, information is but one of the objectives of education; certain attitudes, appreciations,

and skills are as important and perhaps more so than information. There is a need to fill in the gaps of our testing programs with reliable methods of measurement which will more completely evaluate the outcomes of the educative process in terms of all its important objectives.

FRED P. FRUTCHEY

ODELL, C. W. *Educational Tests for Use in High Schools*, Third Revision. Urbana, Illinois: University of Illinois, 1929. 50 pp. (University of Illinois, Bureau of Educational Research, Bulletin No. 53)

The purpose and administration of standardized educational tests for high schools are discussed in the light of their functions, present status, and limitations. Five criteria, which are valuable in selecting tests, are given. A list of references on testing which will be helpful to those engaged in teaching or in instituting a testing program is included.

FRED P. FRUTCHEY

WOODY, CLIFFORD. *Results of the May Testing Programs.* Ann Arbor, Michigan: Bureau of Educational Reference and Research, School of Education, University of Michigan, 1929. 49 pp. (Bureau of Educational Reference and Research, Bulletin No. 126)

WOODY, CLIFFORD. *Results of the May Testing Programs.* Ann Arbor, Michigan: Bureau of Educational Reference and Research, School of Education, University of Michigan, 1929. 46 pp. (Bureau of Educational Reference and Research, Bulletin No. 132)

These reports give the results of the testing programs which were carried on in the elementary schools and high schools of the state of Michigan during the years 1928-29 and 1929-30. The results of the programs are presented in a series of tables. The results for both cities and population groups are reported in terms of grade-levels instead of in terms of raw scores. Standards are given for the different groups of cities, for the state as a whole, and for other comparable groups. The author believes it possible to determine the amount of growth made in a grade during the year by a comparison of the October and the May scores of the same group. The October tests may serve as a pre-test in determining what the students do not know and what should be emphasized during the year.

FRED P. FRUTCHEY

PASADENA. *Educational Accomplishment.* Pasadena, California: Pasadena City Schools, 1929. 54 pp. (Administrative Research Series, 1929, Monograph No. 3, Bureau of Administrative Research, Pasadena City Schools)

The testing and measurement techniques employed by the research staff of the Pasadena schools in evaluating educational accomplishment is set forth in this bulletin. Graphs and tables indicate the holding power of the school and the degree of ageness, over-ageness and under-ageness of the pupils. Nine factors are considered as a basis for ability grouping, but in practice the grouping is largely based on intelligence quotients and educational quotients. The number of exceptional cases not properly classified in this way is significant and points to the necessity for a more detailed study of individual children.

FRED P. FRUTCHEY

Books to Read

KRIEGER, LAURA B. M. *Prediction of Success in Professional Courses for Teachers.* New York: Teachers College, Columbia University, 1930. 77 pp. (Teachers College, Columbia University, Contributions to Education, No. 420)

This is a statistical study which aims to aid educational guidance on the level of the graduate school. The book gives a review of the findings of previous studies to determine the value of psychological examinations as predictive measures. The scores of the four predictive tests and the average grades of academic achievement, the criterion, are converted into T-scores, and the reliability of each of the predictive tests was determined by the half-test method. The partial correlation method was used to develop the regression equation for predictive purposes. The research was further continued to study the discrepancies between academic grades and the predictive tests scores. A comparative study was made of the candidates for the Master's degree and for the Doctor's degree, of the differences among departments of Teachers College, of sex differences, and of age differences. The study suggests the inclusion of other predictive factors and of the success in life-work rather than academic success as the criterion.

FRED P. FRUTCHEY

MCADORY, MARGARET. *The Construction and Validation of an Art Test.* New York: Teachers College, Columbia University, 1929. 35 pp. (Teachers College, Columbia University, Contributions to Education, No. 383)

This dissertation describes the construction and validation of a test in art appreciation. The test includes sixty-one items, each comprising four illustrations of art objects. The pupils are asked to rank the four illustrations in each item in the order of artistic merit. The illustrations have been so chosen that the total test includes variations in line, notan (dark and light), and color.

The ranking of these illustrations by which the test is scored was determined by the consensus of the rankings of one hundred judges who are art dealers, critics, buyers, and advanced students. The items finally included in the test are those in which there was substantial agreement among the judges as to the relative order of merit. The basis for scoring selected, because it had the highest reliability, was to allow one point for each illustration which was ranked in the exact position given by the judges.

In using this test, two criticisms need to be considered. In the first place, it is called a test of "art appreciation." The term "appreciation" has two common implications. It is often used to mean enjoyment. In this sense, a test in "art appreciation" would be a measure of the student's enjoyment of objects possessing artistic merit. This is an important quality to measure, but the test does not provide such a measure. The term "appreciation" is also used to mean the recognition of objects which most nearly meet the accepted standards of art. In this respect the test is an art-appreciation test, but the two meanings of the term are not identical. It is possible for a pupil to recognize a painting by a master as one which is considered artistic, and still the pupil might really prefer to see the pictures in the *Police Gazette.* If one is seeking a test of art enjoyment, this test is not suitable.

The second criticism is in the selection of the method of scoring. The scoring method was chosen because it had the highest reliability. Scoring is a matter of validity as well as reliability. If the proper order of the illustrations is 3–2–4–1, a pupil ranking them 1–3–2–4 would receive no points credit, although he had recognized the descending value of 3–2–4 and had placed in proper relative position all of the illustrations except No. 1. On the other hand, a pupil ranking them 4–2–3–1 would receive one point of credit, although none are in proper relative position except No. 1. To determine the method of scoring such a test is a problem of validity as well as reliability; hence, the scoring method adopted may be criticized. Nevertheless, the test will be useful in providing a measure for the development of pupils' judgment of artistic merit.

R. W. TYLER

MANN, CLAIR V. *Objective Type Tests in Engineering Education.* New York: McGraw-Hill Book Company, 1930. x + 122 pp.

This book gives an account of a study in construction, administration, and evaluation of a series of objective type tests designed for use in college engineering drawing and descriptive geometry. The principal purposes in constructing the tests have been for the measurement of the knowledge of subject-matter, the power to think in terms of it, and the ability to make the desired useful application of the course materials. Two types of tests have been constructed, one being the placement examination for the determination of aptitudes and training for engineering drawing and descriptive geometry and the other, the achievement type which purports to measure classroom accomplishment.

FRED P. FRUTCHEY

MADSEN, I. N. *Educational Measurement in the Elementary Grades.* Yonkers, New York: World Book Company, 1930. x + 294 pp.

This book, which belongs to the Measurement and Adjustment series, edited by L. M. Terman, is concerned with tests for the elementary grades. It is so written that it can be used as a text for most of the beginning courses in educational measurements. A brief history of measurement in education and of individual differences is presented to give the background out of which objective tests have developed. Such procedure as tabulation, as well as measures of central tendency, measures of variability, percentiles, correlation, and other statistics needed by teachers are discussed. Validity and reliability of tests is treated briefly but rather well. The book includes material on intelligence tests, on the assumption that the educational test program is incomplete without it.

The chapter on achievement tests gives a fairly good description of a number of typical tests for all the elementary grades. Included are survey tests, diagnostic tests, and practice tests. The chapter on educational uses of standardized tests is perhaps the most useful chapter of the entire book. The testing program for the entire school year is outlined, emphasizing the necessity of action after the tests have been given and scored. A bibliography is given at the end of each chapter.

HAROLD A. EDGERTON